

PR #2602 完整报告

radixark/miles

glm52_tbench2: default the train set to the 69-task split

合并时间: 2026-08-18 18:07

原文链接: <http://prhub.com.cn/radixark/miles/pull/2602>

执行摘要

- 一句话: glm52_tbench2 默认训练集切到 69 任务, 修复小型任务集训练发散
- 推荐动作: 值得精读, 尤其是 PR body 的诊断叙事: 它展示了从 A/B 实验到 per-token dump 取证、再到机制定位的完整调查链, 并得出一个有迁移价值的设计决策——当小任务集上同源负 advantage 更新反复自放大时, 扩大任务集稀释重复比继续调 knob 更强效。对于从事 RL 训练稳定性、agentic rollout 调优的工程师, 这个案例是很好的参考。源码层面改动很小, 精读重点是论证逻辑而非代码本身。

功能与动机

PR body 给出了明确的 A/B 证据链: 在 69 任务 (July stack, job 1794) 下, 即使其他轴全部使用更差的旋钮 (16384 per-turn cap、131k session、tis-clip-low 0.5), 训练仍保持 100 步健康; 而在 8 任务 (当前 stack, jobs 2350/2403/2423) 下, 每个运行都进入相同发散——trainer log-probs 在巨大截断轮次的共享内容上反复接受负 advantage 更新而漂移, train_rollout_kl 指数增长。Per-token dump 取证显示 8 任务下失败 / 循环任务每 step 重复发射近乎相同的巨大轮次, 同方向更新自放大 (mean llp_diff 0.06 -> 0.88 -> 3.85), 而 8.6 倍更大的任务集恰好稀释这种重复。

实现拆解

本 PR 的变更以配置默认值调整为核心, 分三步拆解:

1. 切换默认训练集文件名: 在 examples/experimental/openenv/glm52_tbench2/run_glm52_744b_a40b_daytona.py 的 ScriptArgs 类中, 将 prompt_data 字段注释与 __post_init__ 的默认值兜底逻辑从 tbench2_train.jsonl (8 任务) 改为 tbench2_train69.jsonl (69 任务)。该字段是 recipe 的入口级配置, __post_init__ 在用户未显式传入 --prompt-data 时生效, 因此这一改动会影响所有沿用默认值启动的 16 节点 GB300 训练。
2. 同步 README 数据准备命令: 在 examples/experimental/openenv/glm52_tbench2/README.md 中, 将 make_tbench2_data.py 的输出文件名从 tbench2_train.jsonl 更新为 tbench2_train69.jsonl, 与源码默认值保持一致, 确保按文档操作生成的 69 任务文件恰好能被 recipe 默认拾取。
3. 测试与配套: 本次改动未新增或修改任何测试文件, 属于纯配置默认值调整, 无 schema、部署或 CI 配套变更。评估集 tbench2_eval.jsonl 的默认逻辑不受影响, 仍使用独立

holdout 任务集。

关键文件：

- `examples/experimental/openenv/glm52_tbench2/run_glm5_2_744b_a40b_daytona.py`（模块训练脚本；类别 `source`；类型 `core-logic`；符号 `ScriptArgs.prompt_data`, `ScriptArgs.post_init`）：GLM-5.2 744B-A40B GB300 训练的入口 `recipe` 脚本，`ScriptArgs` 类的 `prompt_data` 默认值在此定义并在 `__post_init__` 中兜底设置；本次将默认训练集从 8 任务切换为 69 任务，是修复训练发散的核心变更。
- `examples/experimental/openenv/glm52_tbench2/README.md`（模块使用文档；类别 `docs`；类型 `documentation`）：README 中的数据准备命令与源码默认值强绑定，若不同步更新，按文档生成的训练集文件名将无法被 `recipe` 默认拾取，造成文档与行为脱节。

关键符号：`ScriptArgs.post_init`

关键源码片段

`examples/experimental/openenv/glm52_tbench2/run_glm5_2_744b_a40b_daytona.py`

GLM-5.2 744B-A40B GB300 训练的入口 `recipe` 脚本，`ScriptArgs` 类的 `prompt_data` 默认值在此定义并在 `__post_init__` 中兜底设置；本次将默认训练集从 8 任务切换为 69 任务，是修复训练发散的核心变更。

```
# run_glm5_2_744b_a40b_daytona.py（节选：训练集默认值及其兜底逻辑）
```

```
class ScriptArgs(U.ExecuteTrainConfig):
    # OpenEnv / Daytona
    # 默认训练集为 69 任务的 tbench2_train69.jsonl。
    # 依据 PR #2602 的 A/B 证据：8 任务时失败 / 循环任务每 step 重复发射
    # 近乎相同的巨大截断轮次，且每个轮次携带一致负 advantage，
    # 同方向更新在共享内容上自放大（mean llp_diff1 0.06 -> 0.88 -> 3.85），
    # 最终导致 train_rollout_kl 指数增长与生成退化。
    prompt_data: str = "" # default: <data_dir>/tbench2_train69.jsonl
    # ...（其余 OpenEnv/ Daytona 配置字段省略）

    def __post_init__(self):
        # ...（节点数与 DAYTONA_API_KEY 校验省略）
        if not self.prompt_data:
            # 未显式传入 --prompt-data 时，默认落到 69 任务训练集；
            # 评估集仍使用独立 holdout 的 tbench2_eval.jsonl，不受本次变更影响。
            self.prompt_data = f"{self.data_dir}/tbench2_train69.jsonl"
        if self.eval_interval and not self.eval_prompt_data:
            self.eval_prompt_data = f"{self.data_dir}/tbench2_eval.jsonl"
```

评论区精华

本 PR 没有实质性的技术 review 讨论。唯一的 review 来自 `claude[bot]` 的自动化提示（说明本仓库配置为手动 review 模式），以及 `Shi-Dong` 的 `APPROVED`（无评论文字）。由于改动

极小且论证全部承载在 PR body 中，评审者直接通过了该变更；PR body 本身即是完整的技术论证文档。

- 无实质技术讨论，论证集中在 PR body (other): 变更直接通过；PR body 中完整的 A/B 证据链和 per-token dump 取证已构成充分论证。

风险与影响

- 风险：主要风险如下：
- 默认行为变更的隐式影响：任何依赖旧 8 任务默认值的脚本、对比实验或自动化流程，在未显式指定 `prompt_data` 的情况下会静默切换到 69 任务集，可能导致与历史 run 的对比口径不一致。
- 缺少测试防护：本次改动没有对应的测试文件，配置默认值变更无法通过自动化测试拦截回归；未来若有人误将默认值改回 8 任务，不会触发任何测试失败。
- 文档一致性风险：README 中仅更新了数据准备命令，仓库内其他位置（如 launch 脚本注释、docs 目录）若仍引用 `tbench2_train.jsonl`，会产生文档与行为不一致的隐患（本次材料未发现其他引用，但需人工确认）。
- 数据集规模副作用：69 任务集更大，单 step 的数据混合更广，step 级统计指标（如 `train_rollout_kl`、`reward` 均值）与历史 8 任务 run 不可直接对齐，后续解读实验曲线时需注意基线变化。
- 影响：影响范围集中在 GB300 16 节点上运行的 `glm52_tbench2` 训练实验：
- 对用户：所有按默认配置启动的 744B-A40B 训练将自动使用 69 任务训练集，预期消除反复出现的 `train_rollout_kl` 指数发散与生成退化问题，显著提升长跑训练的成功率。
- 对系统：训练集规模扩大 8.6 倍会改变每 step 的 prompt 分布，对 rollout 采样、sandbox 调度和 trainer 数据消费有一定负载影响，但不会触及核心框架代码。
- 对团队：该变更是 GB300 训练稳定性调查链的落地一环，为 staged validation 提供了明确的默认配置；它确立了一个可复用的方法论——用更大的任务集稀释重复负梯度更新，而非继续调优超参。

影响程度：中等偏局部。不涉及核心库代码，但直接影响 16 节点大规模训练实验的有效性。

- 风险标记：训练数据默认值变更，缺少测试覆盖，依赖旧默认值的流程需显式指定

关联脉络

- PR #2588 `glm52_tbench2`: collapse levers and launch fixes from the GB300 16-node smoke runs: PR body 明确说明本 PR complements #2588（折叠配置杠杆与启动修复），两者是同一 GB300 16 节点训练稳定性调查线的连续落地，且无文件行级重叠。
- PR #2587 `openenv`: stop the episode at a length-truncated turn: 该 PR 处理截断轮次立即结束 episode 的问题，与本 PR 分析的巨大截断轮次（giant capped turns）在 agent 行为层面直接相关，共同构成对截断行为的完整处理。
- PR #2591（代码注释中引用的 TODO issue）agent loop cannot see the `max_seq_len` cut: `run_glm5_2_744b_a40b_daytona.py` 中 `max_seq_len = 65536` 的注释引用了 TODO(#2591)，说明 agent 循环无法感知 CP 切分后的序列预算限制，与本 PR 讨论的巨

大轮次截断现象有直接关联。